International Journal of Library
Science and Research (IJLSR)
ISSN(P): 2250-2351; ISSN(E): 2321-0079
Vol. 6, Issue 4, Aug 2016, 1-10
© TJPRC Pvt. Ltd.

TRANS STELLAR
•Journal Publications •Research Consultancy

COMPREHENSIVENESS, DEAD LINKS AND DUPLICACY OF SELECT MAJOR SEARCH ENGINES IN THE FIELD OF LIBRARY AND INFORMATION SCIENCE

PEERZADA MOHAMMAD IQBAL[1], ABDUL MAJEED BABA[2] & AASIM BASHIR[3]

[1]*Professional Assistant, Sher-e-Kashmir University of Agricultural Sciences and Technology, Kashmir, India*

[2] *HEAD, Department of Library and Information Science, University of Kashmir, India*

[3]*Assistant Professor, Department of Computer Science, University of Kashmir, India*

ABSTRACT

This paper presents the results of a research conducted on seven search engines- Google, Bing, Yahoo, Ask, Baidu, Dogpile and DuckDuckgo for comprehensiveness, Dead links and Duplicacy of Select search engines using Library and Information Science related search terms. The search engines are evaluated by taking the first twenty results pertaining to "Dead links" & "Duplicacy". It shows that Google is most comprehensive in its database while DuckDuckGo retrieved the most number of dead links. On the other hand DuckDuckGo and Dogpile show most number of Duplicacy in Library and Information Science search terms. The research also reveals that Baidu retrieved the most scholarly URL's in Library Science field.

KEYWORDS: Search Engine, Dead Links, Comprehensiveness, Duplicacy, Library and Information Science

Received: Jun 16, 2016; **Accepted:** Jun 29, 2016; **Published:** Jul 12, 2016; **Paper Id.:** IJLSRAUG20161

Original Article

INTRODUCTION

The World Wide Web can be used as a quick and direct reference to get any type of information in electronic format all over the world. However, information found on the Web needs to be filtered and may include voluminous misinformation or non relevant information. The user or Internet surfer may not be aware of quality search engines to get information on a topic quickly and may use different search strategies. Finding useful information quickly on the Internet poses a challenge to both the ordinary users and the information professionals. Though the performance of currently available search engines has been improving continuously with powerful search capabilities of various types, the lack of comprehensive coverage, the inability to predict the quality of retrieved results, and the absence of controlled vocabularies make it difficult for users to use search engines effectively. The use of the Internet as an information resource needs to be carefully evaluated as no traditional quality standards or control have been applied to the Web. Librarians need to be able to provide informative recommendations to their clientele regarding the selection of search engines and their effective search strategies.

OBJECTIVES

The following objectives are laid down for the study:

- **To Select Search Engines & Search Terms for the Study**

There are countless numbers of search engines over the internet. Some are active while others are inactive, some are country bound other are global, some are subjective, unilingual, etc while others are general, multilingual etc. Selection of search engines will be based on the following parameters.

- Automatic Indexing.
- Global Coverage.
- Advanced search feature.
- Refine searching in Portable Document Format (PDF).
- Providing gist of information while indexing

Since the scope of the study relates to the field of Library and Information Science. The terms will be selected using classifying schemes from Library and Information Science and List of subject headings. The terms will be further refined to into three categories i.e., Simple, Compound and Complex terms.

- **To Find Out Comprehensiveness of Selected Search Engines**

  The study will estimate total results and scholarly documents retrieved from provided search terms.

- **To Identify Duplicate Results and Dead Links in the Search Engine Results**

Duplicate results will be identified when similar contents appear under different URL's in the retrieved hits for each query and Dead links is identified when the web responds with the information "The page cannot be found" or "The page you are looking for is broken or does not exist".

METHOD

As certified by International Standard Organization there are 230 search engines **(Promote3.com, 2015)** available for searching the web. These search engines are of various types like general search engine, robotic search engine, Meta search engine, directories and specialized search engines. Most users prefer robotic search engines as they allow the users to compose their own quires rather than simply follow pre specified search paths or hierarchy as in case of directories. Moreover, robotic search engines locate data in a similar way i.e., by the use of crawlers or worms. This distinguishing feature differentiates them form web directories like Yahoo! Where collections of links to retrieve URL's are created and maintained by subject experts or by means of some automated indexing process. However some of these services are also include a robot driven search engine facility. But this is not their primary purposes. This due to this feature Yahoo! Was included for the study.

Meta search engine e.g., Dogpile etc don't have their own database. These access the database of many robotic search engines simultaneously. Thus these were included for the study.

Still hundreds of robotic general search engines navigate the web, in order to limit the scope of study after preliminary study, following criteria was laid down for selection of general search engines:-

- Availability of automated indexing
- Global coverage to data.
- Quick response time.
- Availability of filter search mechanism
- Least overlapping.

- Major market holder.

Following two general search engines were selected for the study for meeting all the criteria and being comprehensive in nature.

- Google.
- Baidu.

Since the study relates to the field of Library and Information Science. It was felt to include specialized search engine in the study representing question answer search engine i.e., Ask.com & another specific i.e., Bing.

There being no full-fledged search engine in the field Library and Information Science except many associated with library websites. Among those human powered (DuckDuckGo) after preliminary investigation and feasibility in the study was included in the study. Thus the search engines undertaken for evaluation of study are:-

- Google (General)
- Bing (Specific)
- Yahoo! (Directory)
- Ask (Question Answer Search engine)
- Baidu (Country Specific General Search engine)
- Dogpile (Meta search Engine)
- DuckDuckgo (Human Powered Search Engine).

Selection of Terms

Selection of terms is not directly possible in development and multidimensional field like Library and Information Science. Therefore, classification schemes like DDC (18^{th}) and DDC (22^{nd}) were consulted to understand Broad/Narrow structure of Library and Information Science. It helped to get five terms/Fields i.e.,

- Information System.
- Digital Library.
- Library Automation.
- Library Services.
- Librarianship.

These terms were then browsed in "LC list of subject Headings" which provided many other related terms (RT) and Narrow terms (NT). Further NT and RT attached to each other preferred or standard terms were also browsed which retrieve a large number of Library and Information Science terms. At first instance 140 Library and Information Science related terms were identified.

Some terms occurred more than once and duplication removed. It reduced the number to 100. Later terms were divided into three broad groups under:

- Application.
- Transformation.
- Inter-relation.

"Application" denotes utility of Library and Information science in various fields and about 50 terms came under this group. "Transformation" refers to a method of developing or manufacturing library services into practical market and 30 terms fall under this group. "Inter-relation" means transformation/dependence of one subject onto another and 20 terms came under this group.

Further each category is sub-divided into groups.

"Application" into four i.e., "Reference service", "Informatics", "Information Retrieval" & "Information Sources". "Transformation" into two i.e., "Digitization" & "Consortia". "Inter-relation" into two i.e., "Library Network" & "Information System".

The terms in each group were arranged alphabetically and each term was given a tag. Later 19% of the terms were selected from each group using "Systematic Sampling" (i.e., first item selected randomly and next item after specific intervals). It further reduced the number to 19. Finally the selected terms were classified into three groups under "Simple", "Compound" & "Complex Terms" (Table 1.1). This was done in order to investigate how search engines control and handle simple and phrased terms.

"Simple Terms" containing a single word were submitted to the search engine in the natural form i.e., without punctuating marks. "Compound Terms" consisting of two words were submitted to the search engines in the form of phrases as suggested by respective search engines and "Complex Terms" composed of more than two words or phrases, were sent to the search engine with suitable Boolean operator "AND" & "OR" between the terms to perform special searches.

Table 1.1: Keywords

S. No	Simple Terms	Compound Terms	Complex Terms
1	Catchwork	Bibliometric Classification	Digital Library Open Source Software
2	Citation	Citation Analysis	Health Information System
3	Dublincore	Comparative Librarianship	Library Information System
4	Indexing	Digital Preservation	Library Information Network
5	Manuscript	Electronic Repositories	Multimedia Information Retrieval
6	Plagiarism	Library Automation	
7	Reprints	Semantic web	

Selection of Search Results and Filteration Technique

To evaluate the select search engines top 20 results from each search engine was taken into consideration to determine precision. The assessment of top 20 results is supported by Hawking, Craswell, Bailey & Griffiths, (2001) compared 20 search engines using first top 20 search results comparing 54 topics originated by anonymous searchers for measuring search engine qualities. Tongchim, Sornlertlamvanich & Isahara, (2006) used seven search engines for measuring effectiveness of search engines on Thai queries. Their results calculated from binary relevance judgments of the first 20 returned results, using 56 topics. Latter Egelman, S., Cranor, L., & Chowdhury, A. (2006) conducted a study of quality and quantity of (Platform for privacy preferred project) P3P-encoded policies associated with top-20 search results

from three popular search engines viz., AOL, Google, and Yahoo!. The study examined top 20 search results returned by each search engine to build a P3P-enabled search engine and used it to gather statistics on P3P adoption as well as the privacy landscape of the Internet as a whole. Dirk (2008) Evaluated 5 search engines using first top 20 hits for retrieval effectiveness of web search engines.

Andago, Phoebe & Thanoun, (2010) collected queries from 30 university students and entered these queries into two search engines viz., Google and Hakia. Precision was thereafter calculated using first 20 hits. The 20 results were taken into evaluated for a comparison of precision of Semantic Search Engine against a Keyword Search Engine. Ajayi & Elegbeleye (2014) used first 20 results for performance evaluation of three search engines. The use of first 20 results were thought to be genuine as majority of scholars use first two pages of search hits which by default to many search engines is fixed to 10 hits per page.

The evaluation of first top 20 hits was further backed by a questioner among the scholars of Kashmir University. A total of 100 questioner were distributed among the Doctorial and M. Phil scholar of said university. The aim was to check the result extension at maximum and type of filtration a scholar uses. It was revealed that 84 percent of the scholar prefer first 20 hits (or two pages: a default of 10 results per page), 11 percent prefer first 10 results and five percent prefer more than 20 results. Further it was revealed that scholar's use PDF (portable document format) to filter the results as to get maximum of research article.

RESULTS AND DISCUSSIONS

Seven search engines were taken into consideration in retrieving Comprehensiveness, Dead Links and Repetition of search terms in the field of Library and Information Science. The process of data collection was carried out from 5th of March, 2016 to 10th of May, 2016. A data sheet was taken into consideration for collection and proper scrutiny of data. The data analysed under following three sections and is supported with figures:-

- Comprehensiveness (Versatility of database).
- Duplication (Repeated Results).
- Dead Links (File not found/ Broken link)

- **Comprehensiveness (Versatility of Database)**

Comprehensiveness of a search engine is defined as a measure of total results a search engine indexes from web and includes the same in its database. A search engine is considered to be more effective when retrieved results are higher against a keyword or a discipline from various dimensions. For example to achieve results against "Software" a search engine may retrieve 10,000 hits but on "Library Software" or "Library Automation Software" it may retrieve 1000 to 3000 hits respectively, where as it may retrieve more or less in other search engines. While analyzing comprehensiveness of select search engines the study took into consideration their "Total Results" and "Scholarly Publications" for the select search terms.

Figure 1.1 and **Figure 1.2** elucidates comprehensiveness of 7 major search engines in terms of total number of web results as well as scholarly publications for select terms. It is obvious from the figure that Google retrieved the largest number of sites. It located 66,71,39,600 URLs for simple queries (i.e.,1-7), 11,12,28,000 URLs for compound queries (i.e., 8-14) and 99,75,00,000 URLs for complex queries (15-19). In total it retrieved 1,77,58,67,600 URLs for the nineteen

queries (1-19) and furnished the highest number of results (69,40,00,000) for the complex query 'Health Information System'. Out of all results it shows 1%, 68% and 31% scholarly documents for simple, compound, and complex queries respectively.

Bing located 83,51,40,000 URLs for all the 19 queries (1-19). Out of which 9, 76, 07, 000 hits are for simple queries (1-7). 3,09,83,000 hits for compound queries (8-14) and 70,65,50,000 for complex queries (15-19). The highest number of URLs (63,70,00,000) is found for complex query 'Library Information Network'. It exhibits 1%, 59% and 40% scholarly documents for simple, compound, and complex queries respectively.

Yahoo! Retrieved 89,55,59,000 URLs for all queries out of which 8,57,17,000 URLs for the simple queries (1-7), 3,88,52,000 for compound queries (8-14) and 77,09,90,000 hits for complex queries (15-19). The highest number of results (39,30,00,000) are for the complex query 'Health Information System'. The engine shows 2%, 59% and 39% scholarly documents for simple, compound, and complex queries respectively.

Ask located 6,058 URLs for all the 19 queries (1-19). Out of which 1,640 hits are for simple queries (1-7). 2,410 hits for compound queries (8-14) and 2,008 for complex queries (15-19). The highest number of URLs (730) is found for complex query 'Multimedia Information Retrieval'. It exhibits 3%, 68% and 29% scholarly documents for simple, compound, and complex queries respectively.

Baidu Retrieved 36,62,37,200 URLs for all queries out of which 4,26,84,000 URLs for the simple queries (1-7), 14,68,83,200 for compound queries (8-14) and 17,66,70,000 hits for complex queries (15-19). The highest number of results (10,00,00,000) are for the compound query 'Citation Analysis'. The engine shows 8%, 66% and 26% scholarly documents for simple, compound, and complex queries respectively.

Dogpile located 9,184 URLs for all the 19 queries (1-19). Out of which 3,337 hits are for simple queries (1-7). 3,040 hits for compound queries (8-14) and 2,807 for complex queries (15-19). The highest number of URLs (870) is found for simple query 'Indexing'. It exhibits 6%, 54% and 40% scholarly documents for simple, compound, and complex queries respectively.

DuckDuckGo Retrieved 3,261 URLs for all queries out of which 1,141 URLs for the simple queries (1-7), 1,228 for compound queries (8-14) and 892 hits for complex queries (15-19). The highest number of results (180) are for the multiple queries viz., Citation, Library Information Network and Multimedia Information Retrieval. The engine shows 3%, 66% and 31% scholarly documents for simple, compound, and complex queries respectively.

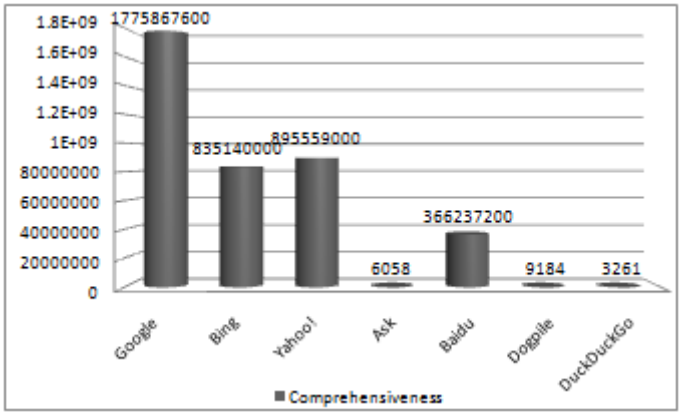

Figure 1.1: Total Comprehensiveness of All Search Engines

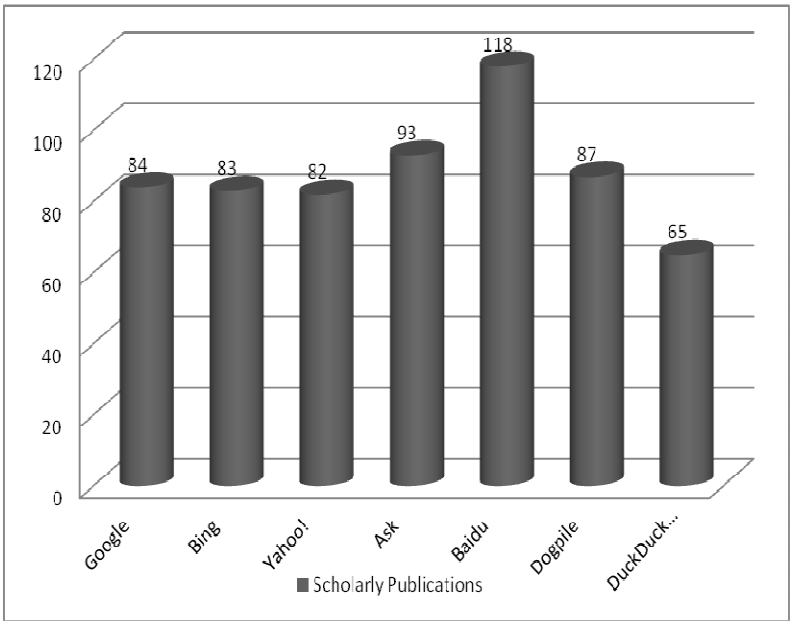

Figure 1.2: Scholarly Publication of All Search Engines In First 20 Hits.

DUPLICACY (REPETITION OF RESULTS)

Many results occur repeatedly under different URLs in the first 20 or first 2 pages (Default of 10 results per page) of search hits of major search engines. Those pages (Scholarly as well as other links) were reviewed for their contents and analysed in Figure 2.1 & Figure 2.2 for all selected queries. It is obvious from the table that DuckDuckGo has 26 hits which are repeated in 2660 results. It means that there is duplication of 0.98% hits among DuckDuckGo search results. Bing results in 20 repeated hits occurring in 2660 results i.e., 0.79% of results occurring twice. Dogpile has shown 0.75% repeated links whereas Yahoo! has 0.71% links occurring twice. Google and Baidu remarkably show 0.38 % of duplicacy and Ask.com has the least occurrence of 0.26% duplicacy among all search engines.

Table 2.1: Duplication among the Evaluated Search Results

Search Engine	Total Results Evaluated	Number of Duplicate Hits
Google	2660	10
Bing	2660	21
Yahoo!	2660	19
Ask	2660	7
Baidu	2660	10
Dogpile	2660	20
DuckDuckGo	2660	26
Total	**18620**	**113**
Average	**2660**	**16.14**

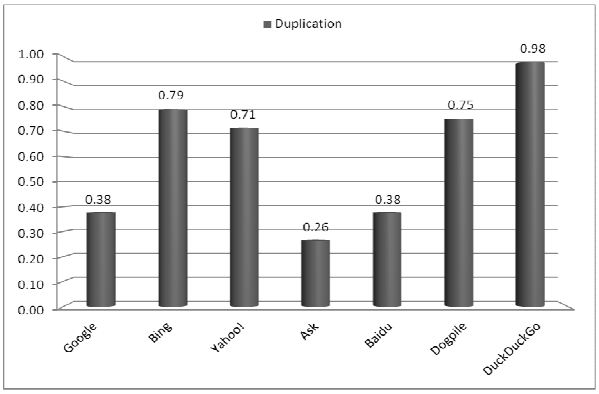

Figure 2.2*:* Percentage of Duplicate Results

DEAD LINKS (FILE NOT FOUND/ BROKEN LINK)

Dead links is defined as a web address having a URL in the result set of a search engine, but page is not displayed on the web. The link is identified when the web responds with the information "The page cannot be found" or "The page you are looking for is broken or does not exist". The links were visited may times under different browsers to avoid cases where the page may be available, but retrievable due to technical errors. All the first twenty hits were reviewed for dead links.

Figure 3.1 and **Figure 3.2** elucidates the number of dead links among the search engine results for all the 19 search queries. It is evident that DuckDuckGo and Dogpile has 8 dead links among the 2660 results evaluated i.e., 0.30% of the documents that do not actually appear on the web although URL exist in the result sets of search engine

Table 3.1: Dead Links in the Select Search Engine

Grand total	Total results Evaluated	Number of Dead Links	Percentage of Dead Links
Google	2660	7	0.26
Bing	2660	4	0.15
Yahoo!	2660	4	0.15
Ask	2660	5	0.19
Baidu	2660	4	0.15
Dogpile	2660	8	0.30
DuckDuckGo	2660	8	0.30

Google retrieved 7 results that have no web link i.e., 0.26% of the hits in Google pages are not accessible.

Ask retrieved 5 hits that have no further link on the web. This percentage of dead links in Ask.com search engines is about 0.19%.

Bing, Yahoo! and Baidu have the lease number of dead links. They retrieved only 4 hits that were not connected to the web, making percentage of dead links in Bing, Yahoo! and Baidu searching to 0.15%.

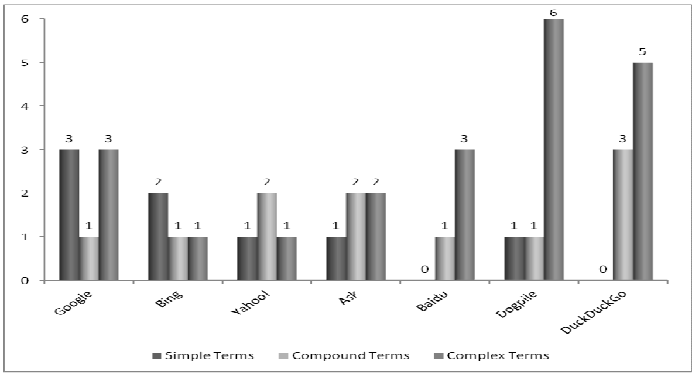

Figure 3.2: No. of Dead Links Retrieved By Given Queries in Select Search Engines

CONCLUSIONS

The immense size and constant transformation of web has made it impossible for search engines to retrieve relevant information. To overcome the problem a variety of search tools ranging from general to subject specific search engines have been developed to provide considerable assistance for finding relevant information on the web. Various studies have been carried out to evaluate search engines for their performance and identify problems associated with them. But due to dynamic nature of search engines each study will not prove final over time.

This study attempted to find the best search engine which is high in comprehensiveness, low in dead links and has least number of repeated documents in the field of library and information science. The study has no inclination towards any search engine as results are purely based on evaluation of the results. The study reveals that those who intend to find high comprehensiveness, low in repetition and least number of dead links in the field of library and information science, Google provides the best opportunity. Yahoo! is the best alternative for getting web resources as its comprehensiveness is also very large with less number of repeated results and least number of dead links. Bing offer good combination of minimal dead links and repeated results along with comprehensiveness. Dogpile being a meta-search engine still lack far behind as low in comprehensiveness, more having dead links and quantum in repetition along with Ask.com and DuckDuckGo.

REFERENCES

1. *Ajayi, O. O., & Elegbeleye, D. M. (2014). Performance Evaluation of Selected Search Engines. Computer Engineering and Intelligent Systems, 5(1), Retrieved from http://www.iiste.org/Journals/index.php/CEIS/article/viewFile/10301/10504*
2. *Andago, M.O., Phoebe, T., & Thanoun, B.A.M. (2010). Evaluation of a Semantic Search Engine against a Keyword Search Engine Using First 20 Precision. International Journal for the Advancement of Science & Arts, 1(2), 55-63. Retrieved from http://www.ucsiuniversity.edu.my/cervie/pdf/ijasa/paperV1N2IT3.pdf*
3. *Dirk, L. (2008). The Retrieval Effectiveness of Web Search Engines: Considering Results Descriptions. Journal of Documentation, 64(6), 915 – 937. DOI: 10.1108/00220410810912451*
4. *Egelman, S., Cranor, L., & Chowdhury, A. (2006). An Analysis of P3P-Enabled Web Sites among Top-20 Search Results. In proceedings of the 8th International Conference on Electronic Commerce (pp. 197 - 207). Retrieved from http://casos.cs.cmu.edu/publications/papers/icec06.pdf*
5. *Hawking, D., Craswell, N., Bailey, P., & Griffiths, K. (2001). Measuring Search Engine Quality. Information Retrieval, 4(1), 33–59. DOI: 10.1023/A:1011468107287*

6. *Promote3.com (2015). Top Search Engine Ranking Search Engine Optimization. IDV International: California. Retrieved from http://www.promote3.com/search-engine-230.htm*

7. *Tongchim, S., Sornlertlamvanich, V., & Isahara, H. (2006). Measuring the Effectiveness of Public Search Engines on Thai Queries. In: Proceedings of The Fifth IASTED International Conference on communications, internet, and information technology. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.1951&rep=rep1&type=pdf*